Smart Data Selection and Reduction for Electric Vehicle Service Analytics

Jennifer Schoch, Philipp Staudt and Thomas Setzer

Karlsruhe Institute of Technology (KIT)
firstname.lastname@kit.edu

Abstract

Battery electric vehicles (BEV) are increasingly used in mobility services such as car-sharing. A severe problem with BEV is battery degradation, leading to a reduction of the already very limited range of a BEV. Analytic models are required to determine the impact of service usage to provide guidance on how to drive and charge and also to support service tasks such as predictive maintenance. However, while the increasing number of sensor data in automotive applications allows for more fine-grained model parameterization and better predictive outcomes, in practical settings the amount of storage and transmission bandwidth is limited by technical and economical considerations. By means of a simulation-based analysis, dynamic user behavior is simulated based on real-world driving profiles parameterized by different driver characteristics and ambient conditions. We find that by using a shrinked subset of variables the required storage can be reduced considerably at low costs in terms of only slightly decreased predictive accuracy.

Keywords

Battery Electric Vehicles; Service Analytics; Service Usage; Data Reduction

1. Introduction

Battery electric vehicles (BEV) are increasingly used in mobility services such as car-sharing. Often, these services are offered and operated by Original Equipment Manufacturers (OEMs) themselves, taking Drive Now or Car2Go as examples. OEMs are seeking to reduce costs, improve quality and customer satisfaction by offering advanced services. Managerial actions are manifold, ranging from guidance and incentive schemes on how to use a mobility service in a way that extends its lifetime (thereby exploiting potentials to offer the service at lower fees) to predictive maintenance to avoid service level degradation or even car breakdowns during service usage.

One primary means of achieving these goals is the exploitation of the vast amount of on-board data gathered from vehicles in the field through telematics or at periodic inspections. Vehicle sensor data is acquired and processed by the respective electronic control unit (ECU) and on-board-diagnostics (OBD) are performed for the sake of vehicle design validation and verification, to identify warranty relevant information and for the detection of system faults. Meeting the requirements for real-time processing, the ECU is an embedded system, which has very limited storage capabilities in an order of magnitude of kB to MB [1]. On the contrary, data loggers that allow for a recording and storage of sensor signals with a high frequency, are limited to the development phases, and therefore rarely represented in series vehicles [2], [3]. Overall, the collection of required sensor signals for the development of new customer services is highly limited by the storage capabilities of today's ECUs used as well as the transmission capacity of telematics.

Hence, to reveal the potentials of smart data analytics, intelligent methods are required to extract the information from sensor data that is most relevant to a respective descriptive or predictive analytical task. In this paper we focus on the collection of data of BEVs in the context of battery degradation.

The propagation of currently available BEVs is mainly impeded by the storage system - the lithium-ion battery - which limits range and leads to long recharging times as well as high costs. Apart from the issues arising with a new BEV, the battery experiences degradation with time and cycling. This manifests in a gradually de-

URI: http://hdl.handle.net/10125/41345
ISBN: 978-0-9981331-0-2

creasing battery capacity and implicates an irreversible reduction of battery capacity, i.e. available range. The progression of battery degradation is highly driven by the user behavior, in terms of driving, charging and environmental factors such as the ambient temperature, as well as the battery management system. Whereas functional dependencies and interactions of degradation relevant variables are not yet fully understood, it is key to make use of the already large amount of BEVs yet in the field to overcome this lack of knowledge. Comprehension of the interplay between dynamic user behavior in a car sharing scenario, is not only crucial for guarantee designs, but also for the development of services such as predictive maintenance, eco driving assistance systems or vehicle to grid (V2G) approaches. Therefore, we provide decision support for OEMs on how to collect sensor data for accurate prediction of system states in terms of capacity fade

The paper is structured as follows. In Section 2 we provide an overview of relevant literature. Subsequently, Section 3 introduces our degradation simulation model and the modeling of user behavior and driving profiles at different levels of detail and data selected. In Section 4 we then provide and discuss the simulation outcomes, in particular the influence of parameter values on degradation. We close with a conclusion and overall recommendation in Section 5 and 6.

2. Related Work

In this Section we will overview services in electric mobility and review work on battery degradation and its main drivers. We will then briefly review approaches to reduce the amount of sensor data.

2.1 Services in the Electric Mobility Sector

In the near future, many vehicles will be transmitting data stored in the ECU on-board by telematics. This development is supported by the EU-guideline for eCall that needs to be fulfilled by 2018 [4] and provides the technical prerequisits for many other data based approaches.

These include predictive maintenance strategies, which aims at forecasting of failure rates of technical devices, guarantee and service design [5]. As a result the occurrence of faults is minimized and consumer satisfaction is increased. Furthermore, location based services benefit from the increasing amout of data and information, such as locating or placement of charging stations, routing, fleet management and car sharing [6].

Literature in the field of smart charging strategies focuses on the flexibility of the EVs storage system while in idle mode, but mostly builds upon data from vehicles with an internal combustion engine. Strategies that aim at balancing the energy grid (vehicle to grid - V2G) [7], [8] or degradation optimized charging strategies [9] will considerably benefit from a large database arising from EV currently in the field.

2.2 Drivers of Li-Ion Battery Degradation

Li-Ion batteries have become the standard storage system for currently available BEVs, due to their high energy density, low self-discharge rate and not exhibiting a memory effect. However, the battery is heavy, costly and furthermore degradation is a severe problem [10].

Battery degradation occurs under both cycling and storage, as cyclic and calendaric aging, respectively [11]. Both types of aging lead to a decrease of the initially available capacity denoted by the state of health (SoH). SoH is typically a relative measure corresponding to the ratio between current capacity and the capacity of a new cell (both at full charge). For automotive applications, the end of life (EoL) for batteries is frequently defined at 80% of the initial capacity ($SoH = 80\%$) [11], [12]. The time and distance covered before the threshold is reached, varies considerably depending on the usage profile. Besides the capacity decline, degradation exhibits an increase of the internal resistance, which affects the power draw capabilities, required e.g. for acceleration. Since the capacity decline is especially challenging for users, needing to deal with the limited range in every day situations, this paper is focused on the capacity fade.

From accelerated aging tests, the main degradation drivers have been analyzed. *Calendaric aging* has been found to be driven by the state of charge (SoC) and temperature (T). Many analyses have revealed a doubling of the degradation when temperature increases by $10°C$. This relationship is usually described by the Arrhenius law ($\sim exp(-\frac{E_a}{RT})$) (with the universal gas constant R and the activation Energy for the capacity fade process E_a) (for example: [12], [13], [14]). Recapitulatory, calendaric aging leads to a monotonically declining capacity with time, while the decline is typically fostered with higher temperatures and higher SoCs [11].

Cyclic aging leads to a monotonic decline of the initially available capacity with the charge throughput (Q), i.e. the accumulated ampere-hours the battery has experienced. The functional relationship is frequently described by a square root function ($\sqrt{Q}$) [13], [14]. A high depth of discharge (DoD) – the SoC range in which cycling occurs – increases degradation rate, while a low DoD around a medium SoC ($\overline{SoC}$) is expected to de-

crease degradation rate [11], [10]. It has been shown by [15] that operation is still possible, but beyond reaching the EoL criterion, degradation rate may increase considerably [12].

Accelerated aging tests constitute a standard method to evaluate battery degradation, but lack to cover the dynamic load that a battery experiences in real world applications. Data from the field allows to overcome this gap of information. However, the potential vast amount of data from BEVs in the field yields to issues with onboard data storage and transmission. Therefore, in the next subsection we will briefly review work on different approaches for the reduction of sensor data.

As aforementioned, to estimate or predict battery degradation, sensor data is required to gather the required parameterization of the respective models. These data needs, however, to be compressed for technical and economical reasons, which impacts the accuracy of a model.

2.3 Sensor Data Acquisition and Reduction

Due to increasing numbers of sensors in different fields such as the automotive industry, industrial production, health sector, mobile devices, fitness and life tracking (quantified self) [16], suitable data acquisition and processing is becoming increasingly relevant in order to make use of the data. However, data reduction is necessary to meet the challenges of energy consumption of sensors at high sampling frequencies, the communications costs that arise when data is transmitted to the base station and the limited storage on embedded systems [17], [3]. Reducing the amount of data can be achieved with different approaches of supervised and unsupervised approaches. Reduction of data (unsupervised) can be achieved with principal component analysis and Fourier- and Wavelet-transformations.

In contrast, if the goal is to explain or predict a particular target variable –in our realm the capacity fade– using the remaining data variables as explanatory features, the nature of the data reduction problem changes. Here, we are in a regression setting where the loss function is solely related to error when approximating or predicting the target variable (supervised reduction of information). Here, for instance, methods to select relevant subsets of sensor-signals are advised, using for example shrinkage methods such as the Lasso regression. Also, a coarser-grained representation of the explanatory variables might be beneficial, given a low increase of predictive error. Filtering data by means of sampling techniques has also been successfully applied in regression settings [17].

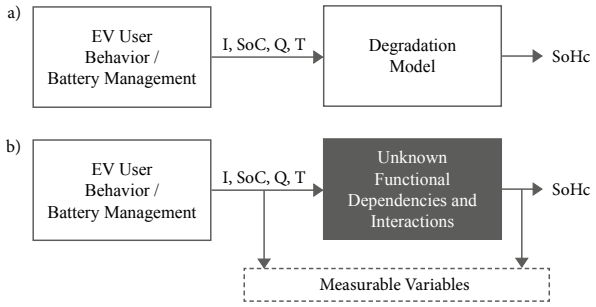

Figure 1: Battery stress factors follow from user behavior and battery management system and the corresponding SoH_C results from the degradation model.

Aiming at a predicting the battery degradation as accurate as possible, under the given restrictions of onboard storage as well as transmission capabilities, transformations and selection of relevant variables needs to be performed and evaluated. To evaluate the trade-off between predictive accuracy and sampled and shrinked subsets of features, we introduce a simulation model based on real-word driving profiles and a degradation model from literature in the following Section.

3. Degradation Simulation Model

Figure 1a indicates how battery stress factors, such as I, SoC (and correspondingly DoD, $\overline{SoC}$), Q and T resulting from a certain user behavior and battery management system strategy are input to a certain degradation model.

The degradation model of the respective type of battery reacts on the stress factors and outputs the respective SoH in terms of capacity (SoH_C). Figure 1b depicts the measurable variables, i.e. stress factors and SoH_C. The degradation model, however is not known in all detail for currently available BEVs.

The following subsections detail the simulation of realistic BEV user behavior, the parameterization of driver types and ambient conditions as well as the degradation model.

3.1 Trip Generation

The simulation of user behavior, throughout the expected battery life of several years, requires a data set of driving profiles of such length with high resolution (acquired by data-loggers). However, to the best of our knowledge, such a dataset is not publicly available. Therefore, our analyses are based on a combination and extension of data from the German mobility panel [18] as well as GPS data logs from the publicly available

Uber Data Set including 25,000 taxi trips within the San Francisco Bay area [19].

The German mobility panel (MOP) is based on the reporting of driving behavior in terms of distance travelled and vehicle location of more than 17.000 households over a period of one week with a resolution of 15 minutes. The mobility panel is separated by the socio-economic background of the participants, while in this paper we focus on the most divers groups of full-time employees and retired. Nine different locations are included in the MOP dataset: *home, work, businesstrip, company trainingcenter, leisure, second home, service, shopping and vacation*.

In order to create driving profiles throughout the lifetime of a BEV battery, the one week MOP driving profiles need to be extended to several years. Therefore, based on the MOP dataset, three empirical distributions are created.

Duration of a stay: Based on all one week MOP profiles, an empirical distribution is created for each 15 minute time slot of a day, differentiated by weekdays and weekends, resulting in $2 \cdot 4 \cdot 24 = 192$ tables for any of nine available locations.

Destination: Similar to the approach for *duration of stay* $2 \cdot 96$ tables are created for weekdays and weekends. Furthermore, the empirical, relative frequencies of occurrences of trips from a start location to an end location are added up to empirical distributions.

Distances: For each start and end location $(9 \cdot 9)$, where start and end location might be identical, relative frequencies are cumulated to empirical distributions.

With the start of the simulation each specific trip is assigned a distance by drawing a random number. That distance remains constant for a given amount of time, typically one year. We choose this design to account for the constancy of many daily distances, for example the trip from home to work or shopping, assumed to be typically similar for a certain period of time. The duration of a stay as well as the next destination are being chosen randomly after each trip, based on the empirical distributions. However, SoC restrictions are being taken into account, when a driving sequence is calculated and it is only allowed to charge the vehicle at defined locations according to the charging strategy (cf. section 3.2).

The driving profile, in terms of velocity, is determined based on the Uber data set. Therefore, GPS logs are transformed to distances, with a resolution of one second. The resulting speed profiles are then clustered based on their specific speed and acceleration levels to create different levels of aggression. Increased maximum speed and an increased gradient of speed (acceleration) correspond to increased aggressiveness.

3.2 Driving Profile Parameterization

The modeling of realistic user behavior and temperature is crucial to solve the aim of feature selection. Therefore, table 1 depicts parameters and values used to generate different driving profiles.

T is coupled to the ambient temperature, but may differ in case of cooling or high current load. In this analysis we assume the ambient temperature to correspond to the temperature on cell level. The employed temperature profiles are based on the year 2015 of the cities of Munich, Madrid and Phoenix, with a resolution of one hour and are repeated annually [20, 21, 22].

The battery current I depends on the driver aggressiveness and topological conditions which lead to different accelerations as well as the chosen charging power, in which fast charging corresponds to high currents. Here, we focus on driver aggressiveness which is clustered in five different groups as described in section 3.1. Charging current is considered constant and rather low, assuming 3.6 kW which corresponds to the power of a standard home socket. High power, fast charging is not considered in this analysis, since to the best of our knowledge no degradation model exists that includes current as a parameter (compare section 3.3).

Implicitly, driver aggressiveness impacts Q, which corresponds to the cumulated Ah-throughput. However, charge throughput is primarily related on the distance travelled.

SoC, DoD and $\overline{SoC}$ depend on the overall driving and charging behavior of the user in terms of distance travelled, energy consumption, timing of trips and charging. Distance travelled is defined by trip generation as described in section 3.1. Furthermore, we differentiate between four different charging strategies. *Just-in-time* charging corresponds to a strategy for charging the BEV as late as possible, whereat all trips need to be feasible with the available SoC. *AFAP* (as fast as possible) charging, corresponds to a maximization of SoC. With *corridor charging* two bounds are defined for the start and end of charging, *lower bound charging* instead only considers a lower bound.

In total, subsequent analysis are based on $\binom{2}{1} \cdot \binom{5}{1} \cdot \binom{4}{1} \cdot \binom{3}{1} = 120$ different combinations of the parameters considered.

Table 1: Parameters and values for driving profile generation.

Parameters	*Values*
driver type	Fulltime; Retired
Aggressiveness cluster	1; 2; 3; 4; 5
Charging strategy	Just-in-Time; AFAP; Corridor; Lower Bound
Ambient temperature	Munich; Madrid; Phoenix

3.3 Degradation Model

This paper is *not* supposed to provide an exact or more detailed understanding of battery degradation. Instead, we provide the prerequisites for subsequent research by identifying a suitable representation of degradation relevant variables, by meeting the constraints of storage and transmission capacities. Furthermore, this paper presents methods on how to transform and process BEV degradation related variables in order to achieve a high predictive accuracy.

In a real world scenario the case of Figure 1b applies. The variables arising from user behavior bounded by the battery management system as well as the resulting SoH_c are measurable, but the underlying degradation model with it's functional dependencies and interactions are unknown. To date no real-world measurements are available due to the novelty of the technology. Therefore, a degradation model from the literature is employed in order to simulate the respective ground truth of SoH_c based on simulations of user behavior.

Representing usage based degradation, a degradation model needs to be found that includes all relevant variables of calendaric (t, T and SoC) and cyclic aging (Q, DoD, $\overline{SoC}$ and I). Several models have been reported in literature, that are based on accelerated aging tests of cells. Whereas all models presented here, include calendaric degradation, the cyclic term does either not include $\overline{SoC}$ [23], [24] or does not include DoD [25], [26]. However, no model, to the best of our knowledge, exists that includes the current I, in terms of C-rate (a C-rate of 1 C corresponds to the current required to fully charge the battery within the time of one hour, e.g. the 1 C rate of a 2 Ah battery equals 2 A).

The degradation model developed by [14] includes all relevant variables except for C-rate, and is therefore found to be most useful to simulate the usage related degradation progress. The model consists of a calendaric (equation (2)) as well as a cyclic component (equation (3)), leading to a monotonically decline of the initially available capacity with $t^{0.75}$ and the square root of Q, respectively. Equation (1) depicts the relationship.

$$Capacity = 1 - \alpha_{cal}(T,v) \cdot t^{0.75} - \beta_{cyc}(\overline{v}, DoD) \cdot \sqrt{Q} \quad (1)$$

$$\alpha_{cal}(T,v) = (7.543 \cdot v - 23.75) \cdot 10^6 e^{-\frac{6976K}{T}} \quad (2)$$

$$\beta_{cyc}(\overline{v}, DoD) = 7.348 \cdot 10^{-3}(\overline{v} - 3.667)^2 + 7.6 \cdot 10^{-4} + 4.081 \cdot 10^{-3} DoD \quad (3)$$

The degradation model is based on cell level, therefore SoC and $\overline{SoC}$ correspond to the cell voltage v and $\overline{v}$, which we assume to be linearly mapped ([0-100%] $\rightarrow$ [3.2 - 4.1 V]). The temperature is measured in Kelvin (K).

The battery capacity deployed in the analyses of [14] are much lower (2.15 Ah) than that of a typical traction battery in a BEV (in this work we assume a battery capacity of 18.8 kWh - Table 2). However, interconnecting many cells in series, results in an overall capacity, meeting the requirements for a traction battery. In total $18800Wh/(2.15Ah \cdot 3.6V) \approx 2430$ cells need to be connected in series to model the considered traction battery of 18.8 kWh. Practically, the battery stress factors are divided by the number of cells.

3.4 Simulated Data Set

The energy required for propulsion results from summing the energy required for acceleration, rolling and air resistance ([10]) and we assume the power drawn from the battery corresponds to the power required to propel the vehicle $P_{bat} = P_{propulsion}$.

Vehicle specific parameters required for deriving the battery current from a driving profile (velocity) include drag coefficient c_w, vehicle frontal area A, vehicle mass m, nominal battery voltage U_{nom} and battery capacity C_{Bat}. Furthermore, constants are required and include air density ρ, rolling resistance coefficient c_r and gravitational constant g. Table 2 depicts the parameters and constants.

$$P_{propulsion} = [F_{acc} + F_{drag} + F_{roll}] \cdot V \quad (4)$$

$$F_{acc} = m \cdot a \quad (5)$$
$$F_{drag} = \frac{\rho}{2} c_w \cdot A \cdot V^2(t)$$
$$F_{roll} = c_r \cdot m \cdot g$$

The battery current finally results from Ohm's law ($I = P/U$).

Table 2: Assumed vehicle specific parameters and constants.

Parameters		*Constants*	
c_w	0.29	ρ	$\rho(T)$ $\frac{kg}{m^3}$
A	2.38 m^2	c_r	0.013
m	1195 kg	g	9.81 $\frac{m}{s^2}$
U_{nom}	360 V		
C_{Bat}	18.8 kWh		

The resulting battery current is derived from trips (chapter 3.1, i.e. velocity and acceleration) and is di-

vided by the number of cells as described in chapter 3.3. The *SoC* results from ampere-hour counting based on charge (positive) and discharge (negative) battery current. Similarly, the charging throughput is derived, employing absolute values for ampere-hour counting. Whereas, the degradation model derived from [14] deploys the *SoC* in terms of the cell voltage v, the *SoC* is assumed to be linearly related to v and mapped from $[0,100]\% \rightarrow [3.2,4.1]$ V, with 3.2 and 4.1 V corresponding to the upper and lower cell voltage bounds, respectively.

$\overline{SoC}$ and *DoD* are derived from *SoC*. However, one cycle is defined such that it contains at least one time slot of driving as well as charging, and starts/ends before the next trip. *DoD* corresponds to the *SoC* delta within one cycle and $\overline{SoC}$ is calculated as the $min(SoC) + DoD/2$ within a cycle.

The procedure of trip generation in each time slot, followed by deriving the battery current, and the calculation of battery degradation is repeated until the EoL criterion of 80% is reached. Cumulating time slots corresponds to the respective battery age t. The battery temperature is assumed to correspond to the ambient temperature (T).

An overview of the simulated dataset of 120 combinations of the parameters charging strategy, drivers occupation, level of aggressiveness and temperature is given in the descriptive analysis of the following Section.

3.5 Descriptive Analysis

On average the lifetime of a car is 10 years and it covers 80,307 km, corresponding to 3,931 Ah of throughput. Comparing the covered distance and the overall battery lifetime at the point of reaching the EoL criterion, Figure 2 depicts considerable differences comparing full-time employees and retired. For each parameter combination that includes 'retired', the covered distance at the same lifetime is in nearly all cases lower than that of 'employees'. For example a lifetime of 10 years leads to approximately 50,000 km covered for 'retired', and approximately 100,000 km for 'employees'. This finding becomes especially interesting when thinking of the guarantee design of currently available BEVs. The guarantee that OEMs currently provide, is expected at least with 5-8 years (Nissan Leaf 24 kWh: 5 years or 100,000 km, www.nissanusa.com; BMW i3 18.8 kWh: 8 years or 100.000 km, www.bmw.com; Tesla Model S 85 kWh: 8 years and no range limitation, www.teslamotors.com).

Most OEMs tailor the guarantee on the battery's age or covered distance, but as can be seen from Figure 2 the variables considerably diverge depending on the driver

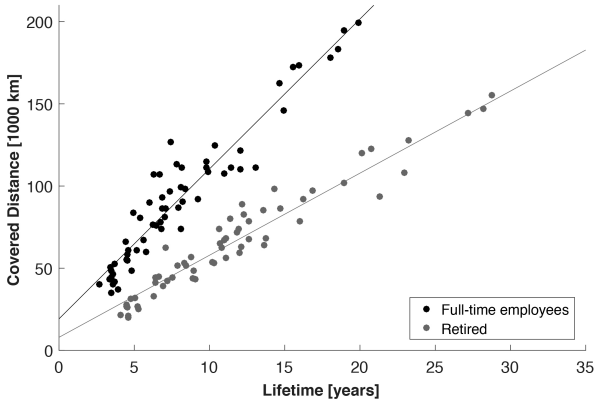

Figure 2: Relationship between the lifetime and the distance covered of each parameter combination at the EoL criterion.

type. From the perspective of a full-employed person, it would be more useful to consider a BEV for purchase that guarantees a certain battery lifetime instead of a distance covered. The contrary applies for retired persons.

Analyzing the influence of each parameter value, two linear regression models with categorical variables have been fitted according to equation 6, for the lifetime $f(EoL) = t(EoL)$ corresponding to coefficients $\beta_0, ..., \beta_4$ and the distance covered $f(EoL) = Distance(EoL)$, corresponding to coefficients $\tilde{\beta}_0, ..., \tilde{\beta}_4$.

$$\begin{aligned} f(EoL) = & \beta_0 + \\ & \beta_1 \cdot DriverType + \\ & \beta_2 \cdot ChargingStrategy + \\ & \beta_3 \cdot AggressivenessCluster + \\ & \beta_4 \cdot T \end{aligned} \tag{6}$$

Table 3: Combination of parameters and the effect on degradation in terms of lifetime in years and distance covered in km.

	$t(EoL)$	$Distance(EoL)$
Coefficient	**Estimate** β	**Estimate** $\tilde{\beta}$
Intercept	19.01^{***}	172180^{***}
ChargingStrategy:AFAP	-10.59^{***}	-68182^{***}
ChargingStrategy:Corridor	-8.54^{***}	-53599^{***}
ChargingStrategy:LowerBound	11.09^{***}	-72955^{***}
AggressivenessCluster:2	-1.69^{**}	-24501^{***}
AggressivenessCluster:3	$-0.15(ns)$	-8026^{**}
AggressivenessCluster:4	$0.62(ns)$	$2200(ns)$
AggressivenessCluster:5	$-0.14(ns)$	$-3507(ns)$
DriverType:Retired	4.0^{***}	-24477^{***}
T:Madrid	-2.74^{***}	-21248^{***}
T:Phoenix	-6.67^{***}	-51304^{***}

The intercept β_0 and $\tilde{\beta}_0$ of both presented regressions with categorical variables corresponds to the reference scenario with *ChargingStrategy*: Just-inTime, *Aggres-*

sivenessCluster: 1, *DriverType*: Fulltime and the temperature T: Munich (Table 3). Coming from the reference scenario with an average lifetime of 19.01 years, battery lifetime is reduced significantly by 10.59 and 8.54 years for AFAP and Corridor charging, respectively. On the contrary, Lower bound charging significantly increases lifetime by 11.09 years. Comparing *AggressivenessClusters* indicates that only cluster 2 yields significant reduction of lifetime of 1.69 years, affecting the lifetime much less than the *ChargingStrategy*. Retired on average lead to an increase in lifetime of 4 years compared to fulltime profiles. Both temperature profiles derived from the ambient temperature in Madrid and Phoenix lead to a decrease of lifetime of 2.74 and 6.67 years, respectively. Looking at the distance covered, any parameter combination deviating from the reference scenario leads to a reduction of the distance throughout the battery's lifetime, as indicated by Table 3. However, coefficients for *AggressivenessClusters* 4 and 5 are non-significant.

In summary, battery lifetime and distance covered before the EoL criterion is reached differs depending on the driver behavior and temperature. This is significantly depending and mainly driven by the charging strategy. In real-world applications it is most likely to observe AFAP charging [27] and currently hardly any smart charging strategy is applied in a large scale. Therefore, it is unlikely to observe large spreads of lifetime as compared to our simulation. However, to the best of our knowledge, no broad results of empirical degradation in EVs have been reported in literature.

After providing descriptive analysis (in-sample), we aim at the evaluation of predictive accuracy (out-of-sample) and evaluate different transformations and shrinkage of features as well as the required data volume in the following.

4. Prediction Model

In this Section transformed, selected and compressed versions of relevant stress factors are evaluated on their predictive accuracy on battery degradation.

As compared to the previous section, not only the time and distance covered to EoL is supposed to be explained, but instead functional dependencies are seeked.

In order to predict the SoH_C progression we differentiate between two approaches. First, the dependent variable corresponds to the monotonously decreasing SoH_C progress. Second, the delta of SoH_C between two subsequent time slots is used as the dependent variable. In the following the first and second approach are called *global* and *delta* model, respectively.

The features created from the trip generation, as summarized in Section 3.1, and the thereupon resulting 40 different battery stress factors are shown in Table 4.

Table 4: Complete feature set. (*) including minimum, maximum, mean, median, 25 and 75% quartiles

Feature	Description	Frequency
t	Battery age	trip
$dist_{total}$	Covered distance	trip
N_{trip}	Total number of trips	trip
f_{trip}	Frequency of trips in trips per year	trip
Q	Charge throughput	trip
$\overline{DoD}$	Depth of discharge per cycle	cycle
$\overline{SoC}$	Average voltage per cycle	cycle
$loc_{beforeTrip}$	Location before trip	trip
$SoC_{beforeTrip}$	SoC before trip	trip
$SoC_{afterTrip}$	SoC after trip	trip
$dist_{trip}$	Length of trip in km	trip
$dist_{cycle}$	Distance covered per cycle in km	cycle
$Q_{perMeter}$	Average consumption per meter	trip
$Q_{perTrip}$	Average consumption per trip	trip
SoC_{rest}	SoC during rest	trip
SoC_{trip}	SoC during driving	trip
SoC_{Δ}	SoC consumption per trip	trip
T_{rest}	Average Temperature during rest (*)	trip
T_{charge}	Average temperature during charging (*)	cycle
V	Average velocity (*)	trip
acc	Average acceleration (*)	trip

In order to evaluate the predictive accuracy of features described in Table 4 linear regression models are employed. A 10-fold cross validation was carried out to evaluate the out-of-sample prediction error. Models are compared based on their normalized root mean squared deviation (NRMSD).

For variable selection and shrinkage the variance inflation factor (VIF), Lasso, Ridge and Elastic Net regression (ENR) is applied. VIF is a measure that identifies collinearity and features are excluded from the model in case the VIF is greater than 10. Lasso is a method for coefficient estimation comparable to ordinary least squares (OLS). However, instead of just minimizing the residual sum of squares as done in OLS, a penalty is put on the sum of L1-norms of coefficients. The penalty is chosen, such that the test error is minimal. Coefficients that are shrunken to zeros correspond to features that are excluded from the model. Ridge regression is comparable to Lasso, and coefficients are shrunken towards zero but will not become exactly zero, and no feature selection is performed. ENR is a combination between Lasso and Ridge Regression and therefore performs feature selection.

Furthermore variable transformation and selection of linear combinations of variables is performed using a combination of principal component analysis and VIF.

Each models predictive accuracy as well as the number of features or dimensions (PCA) is depicted in Table 5. Comparing global regression models, none of the shrinked or in dimensionality reduced models outperform the full model – containing 39 features in total according to Table 4 – in terms of test NRMSD. However,

Global Lasso and *Global Elastic* result in a comparable predictive accuracy compared to the *Global* model, requiring only 24 and 27 out of 39 features, respectively. Similar to the observations for global regression models, *Delta Lasso* and *Delta Elastic* result in low RMSD but do not outperform the *Delta* model including all 40 features.

Delta models are based on the differentiated and log-transformed $SoHc$. NRMSD allows for the comparison of results in different scales, therefore NRMSD allows us to compare global and delta models. However, based on NRMSD delta models overall show better predictive performance as compared to global models. However, *Delta Lasso* and *Delta ENR* models result in NRMSD very close to that of the full model and require only a subset of 32 and 33 variables of the originally 40 variables.

Table 5: Test error (derived from cross validation) for different regression approaches.

Modell	Features/ Dimensions	RMSD	NRMSD
Global	39	0.0097	0.0486
Global VIF	15	0.0131	0.0657
Global Lasso	24	0.0105	0.0521
Global Ridge	39	0.0128	0.0640
Global Elastic	27	0.0105	0.0524
Global PCA	12	0.0164	0.0822
Global Cycle	39	0.0143	0.0531
Global Cycle VIF	15	0.0203	0.1017
Global Cycle Lasso	29	0.0149	0.0748
Global Cycle Ridge	39	0.0177	0.0885
Global Cycle Elastic	30	0.0150	0.0750
Global Cycle PCA	13	0.0268	0.1344
Delta	40	0.3909	0.0418
Delta VIF	23	0.3942	0.0422
Delta Lasso	32	0.3912	0.0418
Delta Ridge	40	0.4025	0.0430
Delta Elastic	33	0.3912	0.0418
Delta PCA	15	0.6609	0.0707

Global models generally are based on features generated per trip. Delta models, however, imply cycle based feature updates. According to the definition of a cycle, serval trips can be included within one cycle and the update frequency is reduced. Therefore, also global models are evaluated by using a cycle based feature update frequency, as depicted in Table 5, but did not outperform delta or global models.

The models presented in Table 5 either include all variables derived from our simulation or are based on shirinked subset of variables or linear combinations of models with reduced dimensionality. However, shrinked models that underwent Lasso regression or variable selection using VIF, do no longer include all variables. These models allow for a reduction of signal recording and are therefore compared to relevant stress factors that were used for simulation in Table 6.

Table 6: Remaining features in each prediction model (*Distance per cycle, only relevant for Delta Models)

Feature	Glob. VIF	Glob. Cyc. VIF	Delta VIF	Glob. Lasso	Glob. Cyc. Lasso	Delta Lasso	Glob. ENR	Glob. Cyc. ENR	Delta ENR
t		x	x	x	x	x	x	x	x
t_{rest}	x	x	x		x	x		x	x
$dist_{total}$		x		x	x	x	x	x	x
N_{trip}						x	x	x	x
f_{trip}			x	x	x	x	x	x	x
Q	x		x	x	x	x	x	x	x
DoD		x	x	x	x	x	x	x	x
$\overline{SoC}$	x				x	x		x	x
$SoC_{beforeTrip}$			x	x	x		x	x	
$SoC_{afterTrip}$				x		x	x		x
$dist_{trip}$	x	x	x	x	x	x	x	x	x
$dist_{cycle}$*	-	-	x	-	-	x	-	-	x
$Q_{perMeter}$			x		x	x		x	x
$Q_{perTrip}$	x		x		x	x		x	x
SoC_{rest}	x	x	x			x			x
SoC_{Δ}					x	x		x	x
T_{rest}	x	x	x	x	x	x	x	x	x
T_{charge}			x	x	x	x	x	x	x
V	x	x	x	x	x	x	x	x	x
acc	x	x	x	x	x	x	x	x	x

Each model that underwent variables selection by using VIF allows to leave out variables related to one or more different, relevant stress factors. The *Delta Lasso* – the model performing best in terms of NRMSD – explicitly includes the all features except for: $SoC_{beforeTrip}$, the mean and 75% quartile of T_{rest}, the 25% and 75% quartiles, median and mean of T_{charge} and the 25% quartile of acc. $SoC_{beforeTrip}$ is highly correlated with $SoC_{afterTrip}$ (0.92), SoC_{rest} (0.85) and SoC_{Trip}, such that the information content is reduced. The statistical moments of T_{rest} and T_{charge} are correlated up to 0.99 such that the selection of moments it not surprising. The 25% quartile of acc does not show an absolute correlation greater than 0.67, but might often be close to zero, explaining the low predictive relevance of this feature. In order to allow for a high predictive accuracy considering a minimum amount of features, it is recommended to focus of the presented, shrinked feature set for degradation prediction purposes.

After evaluating the predictive accuracy of different reduced sets of features, the required data volume needs to be analyzed.

4.1 Data Volume Estimation

By now, we have evaluated the predictive accuracy of different models given the number of predictors or dimensions included in the model. However, we aim at minimizing the required storage that the underlying subset or representation of variables requires, and evaluate the data volume in this Section.

Data reduction is initially achieved by sampling based on trips or cycles. Assuming a sampling of 1

Hz of four relevant signals (SoC, I, T, Q) corresponds to $(4 \cdot 24 \cdot 60 \cdot 60s \cdot 1Hz = 354600)$ data points per day. Having 2.4 and 1.7 trips per day for fulltime employees and retired, respectively, the number of data points per day reduces considerably by factor $354600/(40 \cdot 2.4) = 3600$ and $354600/(40 \cdot 1.7) = 5082$.

We investigate on the models accuracy by predictions in terms of the lifetime in years and distance covered in km at EoL ($SoH = 80\%$). Results are presented in Table 7 using the most promising models of Table 5, considering the models with all features included as well as VIF and Lasso models.

Table 7: Prediction error in lifetime and distance covered

Model	Data volume [kByte/day]	Prediction error age [years]	Prediction error Dist. covered [km]
Global	410	2.48	17,416
Global VIF	146	2.64	17,239
Global Lasso	244	2.49	17,277
Global Cycle	291	1.7	12,077
Global Cycle VIF	109	2.15	15,370
Global Cycle Lasso	164	1.7	12,268
Delta	290	1.72	12,843
Delta VIF	156	1.85	13,516
Delta Lasso	212	1.72	12,842
Parameter model	0	3.7	23,151

Evaluating the simplest model as a benchmark, a regression is performed based on the parameter configuration according to Table 1, indicated by *Parameter model* in Table 7. Throughout the battery lifetime, one constant combination of parameters needs to be derived from driving and charging style and the ambient temperature conditions. Therefore, the required data volume is nearly zero. Any other model, indicated in Table 7, requires considerably larger data volume due to trip or cycle based variable updates. Comparing the predictive accuracy of EoL prediction in terms of lifetime and distance covered, the cycle based, shrinked global models *Global Lasso Cycle* yields the best predictive accuracy – with an average prediction error of 1.7 years and 12,268 km – under minimal data volume of 164 kB per day. As indicated by Table 6, *Global Lasso Cycle* model omits features related to *SoC* which therefore reduces the data volume. The required data volume is well in line with the storage capabilities of a standard ECU for battery management systems, laying in an order of magnitude of kB to MB. Similar results can be achieved by applying the *Delta Lasso* model.

5. Discussion

A simulation of battery degradation has been developed, that considers dynamic user behavior. Based thereupon, we are able to derive implications for battery BEV guarantee design from an OEMs point of view and guarantee (corresponding to BEV) choice from an users point of view, that may differ considerably depending on the driving habits of users. Furthermore, different models have been evaluated based on their predictive accuracy and required storage.

We found that Lasso regression models perform best – compared to dimensionality reduction using PCA and feature selection using VIF – in order to select features with a high predictive accuracy. Moreover, Lasso regression models allow for considerable storage reductions. A higher predictive accuracy can be achieved based on *Delta* models as compared to *Global* models. Resulting subsets of features can be stored onboard a standard ECU assuming daily submission through telematics.

Our analysis currently is simulation based, and can be enhanced through real-world measurements of degradation related signals. Different real-world degradation effects, such as cell inhomogeneities or capacity regeneration has not been considered in this work, but may change the observed degradation process.

6. Conclusion

Using analytical models we have derived a reduced set of features that allows for an accurate prediction of battery degradation in BEVs based on standard equipment. This allows for efficient data acquisition in a fleet of BEVs for example of a car sharing service provider, assuming daily data transmission to a home station through telematics.

Such a resulting database allows for detailed analysis of BEV user behavior and the related battery degradation. Using prescriptive analytics, optimal behavior can be recommended to the user, which will increase the overall efficiency of BEVs including battery lifetime as well as the available range. Car sharing providers may use the insights to map different users, depending on their driving and charging behavior, to the best suited type of BEV. The location of newly build charging station can be optimized based on data gathered from a fleet of BEVs.

From an OEMs point of view, the data allows accurate predictions of the time to EoL and the development of predictive maintenance approaches. Accurate models will result in greater customer satisfaction and therefore increase the retention. It will also cause customers to use the OEMs proprietary service garages and increase revenue.

7. References

[1] W. Sung and C. B. Shin, "Electrochemical model of a lithium-ion battery implemented into an automotive battery management system," *Computers & Chemical Engineering*, vol. 76, pp. 87–97, 2015.

[2] Y. Zhang, G. W. Gantt, M. J. Rychlinski, R. M. Edwards, J. J. Correia, and C. E. Wolf, "Connected vehicle diagnostics and prognostics, concept, and initial practice," *IEEE Transactions on Reliability*, vol. 2, no. 58, pp. 286–294, 2009.

[3] R. Prytz, "Machine learning methods for vehicle predictive maintenance using off-board and on-board data," 2014.

[4] EU, "Richtlinie 2007/46/eg," 2007.

[5] D. Liu, J. Pang, J. Zhou, Y. Peng, and M. Pecht, "Prognostics for state of health estimation of lithium-ion batteries based on combination gaussian process functional regression," *Microelectronics Reliability*, vol. 53, no. 6, pp. 832–839, 2013.

[6] M. Matzner, F. Chasin, M. von Hoffen, F. Plenter, *et al.*, "Designing a peer-to-peer sharing service as fuel for the development of the electric vehicle charging infrastructure," in *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pp. 1587–1595, IEEE, 2016.

[7] C. M. Flath, S. Gottwalt, and J. P. Ilg, "A revenue management approach for efficient electric vehicle charging coordination," in *System Science (HICSS), 2012 45th Hawaii International Conference on*, pp. 1888–1896, IEEE, 2012.

[8] A. Schuller, C. M. Flath, and S. Gottwalt, "Quantifying load flexibility of electric vehicles for renewable energy integration," *Applied Energy*, vol. 151, pp. 335–344, 2015.

[9] J. Schoch, "Modeling of battery life optimal charging strategies based on empirical mobility data," *it-Information Technology*, vol. 58, no. 1, pp. 22–28, 2016.

[10] D. Linden and T. B. Reddy, "Handbook of batteries," 2011.

[11] A. Jossen and W. Weydanz, *Moderne Akkumulatoren richtig einsetzten*. Reichardt Verlag, 2006.

[12] R. Spotnitz, "Simulation of capacity fade in li-ion batteries," *Jounal of Power Sources*, pp. 72–80, 2003.

[13] S. Kaebitz, J. B. Gerschler, M. Ecker, Y. Yurdagel, B. Emmermacher, D. Andre, T. Mitsch, and D. U. Sauer, "Cycle and calendar life study of graphite linmncoo li-ion high energy system. part a: Full cell characterization," *Journal of Power Sources*, 2013.

[14] J. Schmalstieg, S. Käbitz, M. Ecker, and D. U. Sauer, "A holistic aging model for li (nimnco) o 2 based 18650 lithium-ion batteries," *Journal of Power Sources*, vol. 257, pp. 325–334, 2014.

[15] S. Saxena, C. Le Floch, J. MacDonald, and S. Moura, "Quantifying ev battery end-of-life through analysis of travel needs with vehicle powertrain models," *Journal of Power Sources*, vol. 282, pp. 265–276, 2015.

[16] M. Alhonsuo, L. Virtanen, J. Rantakari, A. Colley, T. Koivum, *et al.*, "Mydata approach for personal health–a service design case for young athletes," in *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pp. 3493–3502, IEEE, 2016.

[17] C. C. Aggarwal, *Managing and Mining Sensor Data*. Springer Science & Business Media, 2013.

[18] BMVBS, "German mobility panel (deutsches mobilitätspanel), panelauswertung 2007," *Deutsches Bundesministerium für Verkehr, Bau und Stadtentwicklung*, no. [Online]. Available: http://mobilitaetspanel.ifv.uni-karlsruhe.de, 2008.

[19] U. T. Inc., "Uber gps analysis," 2013.

[20] D. Wetterdienst, "Historische stündliche lufttemperatur station id 3379," 2014.

[21] TuTiempo.net, "El tiempo en madrid," 2014.

[22] W. Underground, "Weather history for kphx," 2014.

[23] A. Marongiu, M. Roscher, and D. U. Sauer, "Influence of the vehicle-to-grid strategy on the aging behavior of lithium battery electric vehicles," *Applied Energy*, vol. 137, pp. 899–912, 2015.

[24] E. Sarasketa-Zabala, E. Martinez-Laserna, M. Berecibar, I. Gandiaga, L. Rodriguez-Martinez, and I. Villarreal, "Realistic lifetime prediction approach for li-ion batteries," *Applied Energy*, vol. 162, pp. 839–852, 2016.

[25] A. Cordoba-Arenas, S. Onori, Y. Guezennec, and G. Rizzoni, "Capacity and power fade cycle-life model for plug-in hybrid electric vehicle lithium-ion battery cells containing blended spinel and layered-oxide positive electrodes," *Journal of Power Sources*, vol. 278, pp. 473–483, 2015.

[26] M. Ecker, J. B. Gerschler, J. Vogel, S. Käbitz, F. Hust, P. Dechent, and D. U. Sauer, "Development of a lifetime prediction model for lithium-ion batteries based on extended accelerated aging test data," *Journal of Power Sources*, vol. 215, pp. 248–257, 2012.

[27] C. C. Rolim, G. N. Gonçalves, T. L. Farias, and Ó. Rodrigues, "Impacts of electric vehicle adoption on driver behavior and environmental performance," *Procedia-Social and Behavioral Sciences*, vol. 54, pp. 706–715, 2012.